TILBURG UNIVERSITY

Tilburg University

A new theory of retailing costs

Nooteboom, B.

Published in:
European Economic Review

Publication date:
1982

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Nooteboom, B. (1982). A new theory of retailing costs. *European Economic Review*, *17*, 163-186.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 06. May. 2023

European Economic Review 17 (1982) 163–186. North-Holland Publishing Company

A NEW THEORY OF RETAILING COSTS*

Bart NOOTEBOOM

Research Institute for Small and Medium-Sized Business, 2509 JE The Hague, The Netherlands

Received April 1980, final version received May 1981

In cross-section studies of shops of the same type, we find that costs in general, and labour in particular, depend on the sales size per shop according to a linear function with a positive intercept. This paper provides some empirical evidence and a theoretical explanation based on queuing theory.

1. Linear cost curves

Retailing does not produce a physical product that can be stocked, but offers a service capacity to be used at the discretion of customers. As a result, a central problem of efficiency is the utilization of capacity in the face of the stochastic nature of customers' arrivals and the different levels of demand during different hours of the day, days of the week and seasons in the year. Also, we might expect non-homogeneous cost functions, because the service capacity of a shop does not approach zero together with the level of sales: there is a minimum or 'threshold' capacity equal to one shop attendant (per point of sale) during opening time, no matter how low the utilization of that capacity is.

In empirical studies of retailing we consistently find linear non-homogeneous relationships between annual costs and annual sales per shop, in any given year, for shops that belong to the same shoptype, which is defined as follows:

Shoptype = class of shops that are homogeneous with respect to assortment composition, service level, own production[1] and mode of supply to the shop.[2]

*This study is derived from my doctoral dissertation defended at the Erasmus University Rotterdam in May 1980, with Professor J. Koerts as supervisor and Professor W.H. Somermeyer as referee. I am indebted to Professor Koerts, Professor J.B.D. Derksen and an anonymous referee for their comments on earlier versions of this paper, which greatly helped me to improve its presentation. I conducted the studies while employed by the Research Institute for Small and Medium-Sized Business in The Netherlands. The results were first reported in a research memorandum in 1977.

[1]Bread baking, pastry-cooking, own slaughter and sausage production, tailoring, repairs, etc.

[2]Chain store, voluntary chain, cooperative, rack-jobbing, leasing, etc.

0165–7410/82/0000–0000/$02.75 © 1982 North-Holland

Later we will give a more precise specification of these homogeneity conditions.

The cost functions are as follows:

$$K = \varphi_0 + \varphi_1 Q, \qquad \varphi_0 > 0, \quad \varphi_1 > 0, \tag{1}$$

where K = annual costs (per shop) and Q = annual sales (per shop).

The relationship is especially pronounced for the volume of labour,

$$a = \alpha_0 + \alpha_1 Q, \qquad \alpha_0 > 0, \quad \alpha_1 > 0, \tag{2}$$

where a = labour volume (expressed in annual total of labour hours or number of persons engaged, in full-time equivalents).

To give some impression of the empirical evidence, we present three graphs, in figs. 1a–c, for independent supermarkets in 1975, independent butchers in 1968, and independent mixed clothes shops in 1975, all in The Netherlands.[3] The justification for establishing separate cost–sales relationships for individual cost items is that the opportunities for substitution between notably labour and shopspace are of little or no

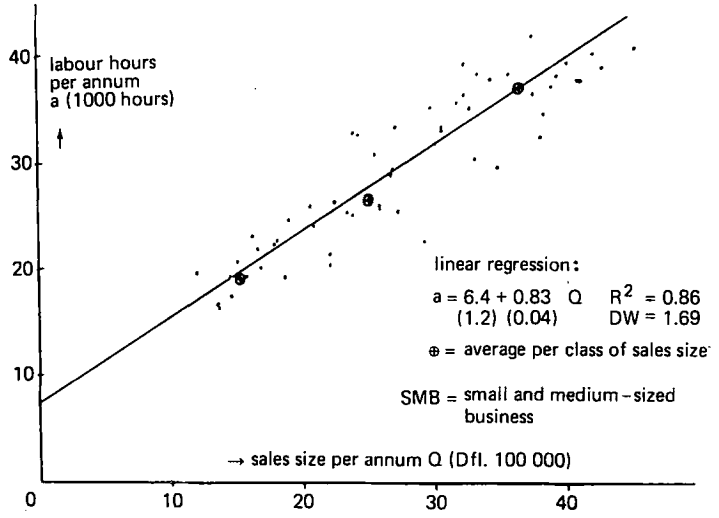

Fig. 1a. SMB supermarkets, 1975.

[3] *Source:* Inter-firm comparisons, Research Institute for Small and Medium-Sized Business, The Hague, The Netherlands.

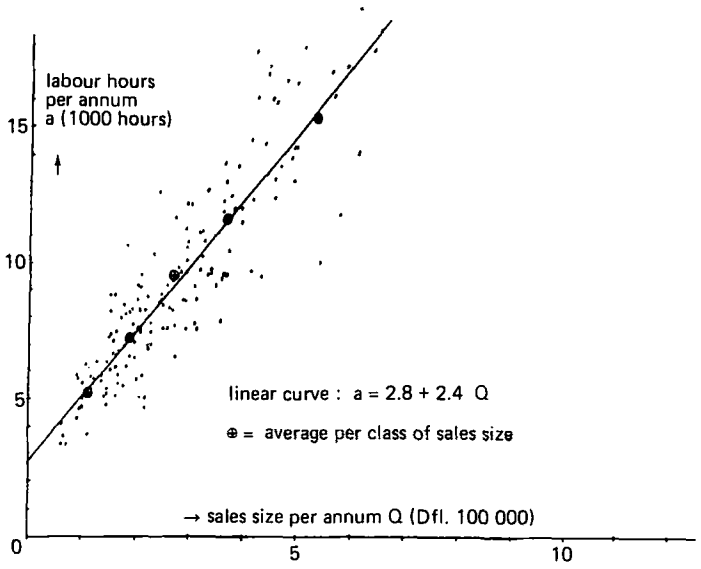

Fig. 1b. SMB butchers, 1968.

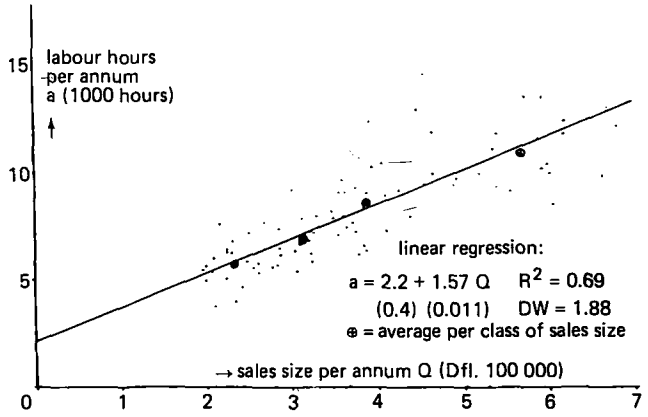

Fig. 1c. SMB mixed textiles, 1975.

importance (under assumptions of homogeneity with respect to assortment, service, own production and mode of supply), as has been noted before by other students of retailing.[4] In the present paper we will discuss only the relationship between annual labour volume and annual sales per shop.

Empirically, the linear cost curve applies in a cross-sectional study of different shops of the same type, in a given year. In such an analysis the sales

[4]Cf. McClelland (1966), Holdren (1960).

size and the capacity of labour and shopspace are variable, and in that sense the curve can be seen as a long-term cost curve for the individual shop (but without time dimension; with an instantaneous adjustment of capacity to a change of sales). The linear cost curve implies an asymptotic increase (in the cross-sectional sense) of the average efficiency of labour (Q/a) with an increase of sales size (Q), as follows: From $a=\alpha_0+\alpha_1 Q$ we find

$$\frac{Q}{a}=\frac{Q}{\alpha_0+\alpha_1 Q} \quad \text{with} \quad \lim_{Q\to\infty}\frac{Q}{a}=\frac{1}{\alpha_1}. \tag{3}$$

This effect of scale is illustrated in fig. 2. This result is in agreement with the phenomenon, noted before by a number of students of retailing,[5] that the economies of scale are very pronounced at the lower end of the scale, while at the higher end of the scale they tend to fade away.

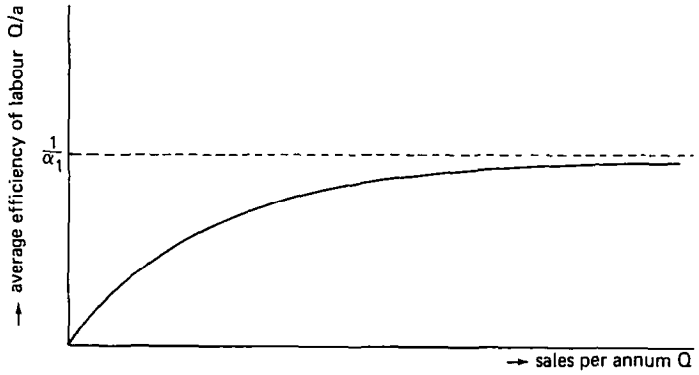

Fig. 2. Economy of scale.

In spite of considerable empirical evidence in favour of the linear cost curve and its implications for the effect of scale, one wants a theoretical explanation and justification of the model in order to adopt it with more confidence. Thus we should be able to:

— explain why the cost curve has a positive intercept (α_0);
— justify a continuous cost curve, in spite of the fact that shop attendants are indivisible, while small shops engage only few attendants;
— explain why the curve is linear.

[5]Twentieth Century Fund (1939), Hall, Knapp and Winsten (1961), McClelland (1966), Pickering (1972), Arndt and Olsen (1975).

2. Explanation

Before we proceed to the formal treatment required for mathematical proofs, we first give a more informal overall view of our line of argument. As indicated before, the cost curve applies to shops with different annual sales sizes, but with the same assortment composition, service level, own production and mode of supply to the shop. More specifically, we assume that the shops:

— have the same number of departments, and the same annual opening times (a line of check-out points in a self-service store counts as one 'department' in the technical sense, in our theory, of a unit which requires at least one person to operate);
— provide the same average service time per guilder sales;
— strive for the same target with respect to the ratio between average waiting time and average service time per customer;
— require the same amount of labour per guilder sales for product handling or, more precisely, for activities that do not arise from direct dealings with customers, which we call 'Pre- and Post-Purchase Activities' or 'PPA' in short (physical distribution, breaking bulk, packaging, storage, price-labelling, display, own production, stock-taking, administration, cleaning, etc.).

We will take into account the fact that a certain proportion of idle service capacity in between customer arrivals can be used for PPA, while this proportion decreases as the proportion of idle capacity decreases, because then the average duration of idle periods, and hence their usefulness, decreases.

There is a straightforward theoretical justification of the intercept α_0. At a given opening time per annum, one must have at least one person (and the requisite shopspace) available during that opening time, to achieve any level of sales at all, no matter how much of the time that person is not occupied effectively. Thus the theoretical value of the intercept is the annual opening time,[6] or more generally, to allow for department stores and large supermarkets (with 'shops in the shop'): there is a 'threshold labour' equal to the sum of the annual opening times of all the separately staffed departments per shop.

A crucial characteristic of retailing is that for any given shop there are different time periods, in the course of a year, with different levels of demand (rate of customer arrivals), depending on the season, the day of the week and the hour of the day. A major instrument of efficiency is the use of part-time labour to adjust service capacity to the different levels of demand in different periods.

[6]In The Netherlands, annual opening time is about 2500 hours.

For a time period with a constant intensity of demand (expected rate of customer arrivals), the relationship between service capacity and sales, in a single department, is a step function as illustrated in fig. 3. In the construction of this figure we used the results of queuing theory which we will develop later.[7] We use the concept of relative waiting time w, defined by

$$w = W/\sigma, \tag{4}$$

where W = average waiting time per customer and σ = average service time per customer.

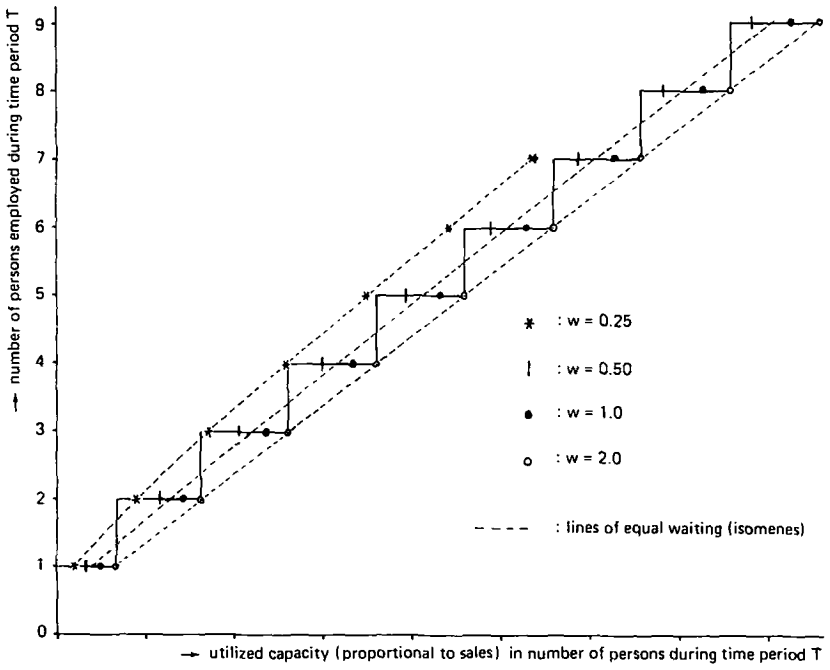

Fig. 3. Labour capacity and waiting time at increasing levels of sales during a period T.

In fig. 3 we assume the following policy with respect to the waiting time imposed on customers: aiming for an average waiting time about equal to average service time ($w \approx 1.0$), the shopkeepers increase their service capacity as the level of sales and hence the mean arrival rate of customers increases, when the average waiting time is twice the average service time ($w = 2.0$). After the addition of one attendant to the service capacity, the average

[7] See table 1.

waiting time falls below the target ($w<1.0$). As the level of sales increases further, the average waiting time lengthens up to the target and then exceeds it up to twice the average service time, at which point yet an additional attendant is added, and so on.

According to the assumption of homogeneity with respect to service, the different shopkeepers in a given shoptype strive for the same target for the waiting time relative to service time (w), with part-time labour as an instrument to adapt service capacity to different intensities of demand in different periods. For any given shop the actual average waiting time will fall short of the target in some periods and will exceed it in other periods. On an annual basis these deviations will tend to offset each other. Summing up: on an annual basis the discontinuity is much reduced, because although attendants are not divisible in the physical sense, labour is divisible in time (part-time labour).[8] We now take a continuous approximation in the sense that we consider continuous curves connecting points with equal values of w, which we call lines of equal waiting or, to coin a name, 'isomenes'.[9] With the approximation in the form of continuous lines of equal waiting, we now propose that on an annual basis the cost curve is as illustrated in fig. 4. Starting from the minimum of threshold labour (α_0, equal to annual opening time, for a single department), the cost curve proceeds along segment $H(Q)$ until it reaches the isomene (line of equal waiting) $G(Q)$ associated with the relative waiting time (w) which is the standard of service in the shoptype in question.

The segment $H(Q)$ will not in general be horizontal (the dotted line in fig. 4), because even for the smaller shop additional capacity above the threshold

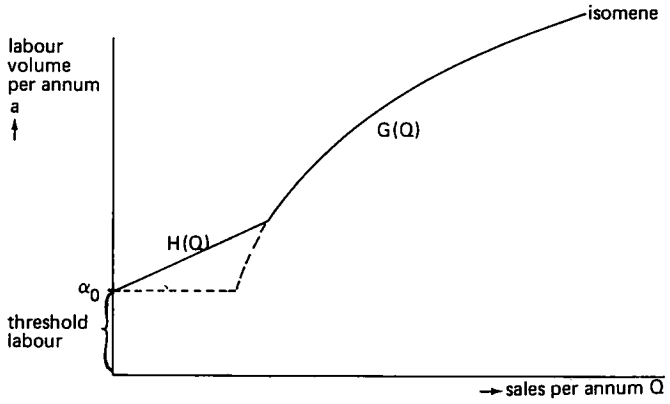

Fig. 4. Cost curve.

[8]For the shopspace the reverse applies: divisibility in space but not in time (no part-time shopspace; except for the interesting phenomenon of the market stall).

[9]Derived from μενειν, the Greek word for waiting.

is generally required during some periods of the year (Saturday morning or afternoon, days before Christmas), while threshold capacity is underutilized during other periods.

The slope of the isomene $G(Q)$ decreases with an increasing sales size Q, because the proportion of idle service capacity in between customer arrivals decreases as the number of attendants increases to keep up with the higher sales level (at a constant value of w and of the average service time per guilder sales). However, the slope of $G(Q)$ decreases at a decreasing rate: the gains of efficiency due to expansion are larger at small sizes than at large ones. In fact, we shall show that the isomene has an asymptote, of the following type:

$$\text{asymptote } y = \beta_0 + \beta_1 Q. \tag{5}$$

We can also prove, on the basis of queuing theory, that for a certain (plausible) range of values of w the intercept of the asymptote (β_0) has the following value:

$$\beta_0 \approx 0.775(1/w)^{5/6}\alpha_0, \tag{6}$$

where α_0 is again the annual opening time.

From (6) it follows that for $w \approx 0.75$ we have $\beta_0 \approx \alpha_0$. In other words: the asymptote then has the same intercept as the cost curve, as illustrated in fig. 5. In this situation we can take the asymptote as a linear approximation of the cost curve. We grant that even under the present rather stringent assumptions (strict homogeneity, use of part-time labour, $w=0.75$) the asymptote does not coincide analytically with the cost curve. We do claim that it yields a satisfactory approximation, as a basis for further research, where we relax the various assumptions, especially those concerning homogeneity. The empirical results (as illustrated in fig. 1) show that the

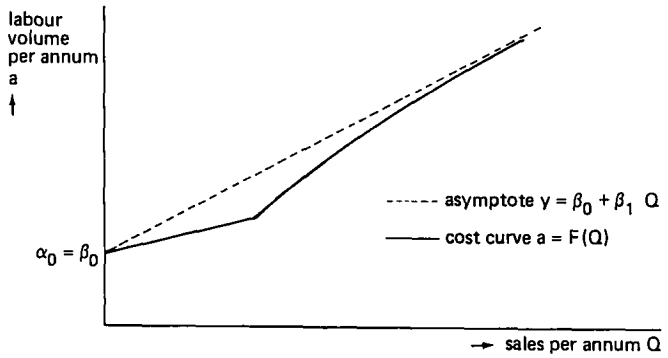

Fig. 5. Linear approximation.

simple linear model performs rather well even in spite of the inhomogeneities and discontinuities that arise in actual observations.

We can also prove that the slope of the asymptote (β_1) has the following value:

$$\beta_1 = \gamma_1 + \gamma_2, \tag{7}$$

where γ_1 = average service time/average sales (per customer) and γ_2 = average labour for PPA (product handling) per guilder sales.

The results, expressed in (6) and (7), apply to shops with a single department. There is a straightforward generalization to shoptypes with more than one department per shop (additional service counters for specialized product groups). If we again assume that only the sales size varies and the shops are otherwise completely homogeneous, and that the average waiting time is about equal to the average service time, the annual cost curve is approximated by an asymptote with intercept β_0 and slope β_1, where

$$\beta_0 = \sum_{i=1}^{n} d_i, \tag{8}$$

where d_i = annual opening time of the ith department and n = the number of departments, and

$$\beta_1 = \sum_{i=1}^{n} s_i(\gamma_{1i} + \gamma_{2i}), \tag{9}$$

where s_i = the share in annual sales, γ_{1i} = average service time/average sales per customer, and γ_{2i} = average labour for product handling, per guilder sales, of the ith department.

It should be noted that with the proper training it may be possible that a single attendant runs two different service counters during quiet hours, in which case the two counters are to be considered as a single department, in the technical sense of our theory. We should also note that for larger shops with several departments the sales-independent share of labour capacity will tend to be increased with one or more supervisors, security agents, hostesses, etc., which then also count as 'departments' in the technical sense of our theory.

After this outline of our theory and our deduction of the linear cost curve, we now proceed to a more formal treatment.

3. Model and proof

In the previous section, our explanation of the linear cost curve ultimately

depends on the following theorem:

Theorem 1

$$a=F(Q) \quad \textit{has an asymptote} \quad y=\beta_0+\beta_1 Q, \tag{10}$$

$$\textit{if} \quad w=0.75 \quad \textit{then} \quad \beta_0=T, \tag{11}$$

$$\beta_1=\gamma_1+\gamma_2, \qquad \gamma_1=\sigma/tr, \tag{12}$$

where

(I) $a=F(Q)$ expresses labour volume (a) as a function of sales (Q) over a period during which the expected rate of customer arrivals does not vary.

(II) T is the length of the period.

(III) γ_1 is the average service time per guilder sales; σ is the average service time per customer; tr is the average sales per customer ('transaction size').

(IV) γ_2 is the average labour time for PPA (product handling), per guilder sales.

We now proceed to prove this theorem on the basis of a mathematical model of store operation, with the following assumptions and definitions:

Assumptions for the application of queuing theory

(V) The arrival rate of customers is assumed to obey a Poisson distribution with expected arrival rate λ (in customers per unit of time).

(VI) The time taken to serve a customer is assumed to obey a negative exponential distribution, with average service time σ.

(VII) We assume that the department is run on a 'first come, first serve' basis; that different service channels operate in parallel; that there is no premature departure of impatient customers in relation to queue length; that the service time is independent of the arrival rate.

(VIII) S is the number of attendants available for serving customers during the time period considered (=number of service channels).

(IX) ρ is the 'utilization factor', defined as the average proportion of service capacity (S) that is actually utilized for serving customers: per unit of time, the amount of labour required to provide the expected service time (σ) for the expected number of customers (λ) is $\lambda\sigma$, so that we have $\rho=\lambda\sigma/S$.

Homogeneity assumptions

(X) The relative waiting time (w) does not depend on sales size (Q).

(XI) The average service time per guilder sales (γ_1) does not depend on sales size (Q).

(XII) The average labour for PPA per guilder sales (γ_2) does not depend on sales size (Q).

Assumption concerning the use of idle service capacity to perform PPA

(XIII) We define ξ as the proportion of idle service capacity that can be used to perform PPA. We assume that it is a function of the utilization factor ρ, as follows: at a low level of utilization of the service capacity, there are long intervals of idle capacity which can be used for a variety of other activities (PPA). It seems plausible, however, that as the utilization of the service capacity (ρ) increases, the intervals of idle capacity become shorter, yielding progressively fewer opportunities for utilization in PPA, so that the useful proportion of idle service capacity (ξ) becomes virtually zero before a 100% utilization of service is reached. It seems plausible to adopt a relationship between ξ and ρ as illustrated in fig. 6.

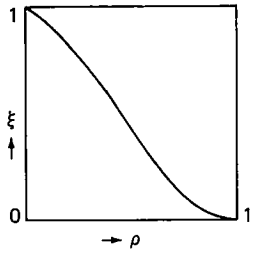

Fig. 6. Utilization of idle service capacity.

In order to represent this, we adopt the following model:

$$\xi=\frac{d(1-\rho)^n}{d-1+(1-\rho)^n}, \tag{13}$$

where d and n are parameters, for which

$$\frac{2n}{n+1}<d<\frac{2n}{n-1}, \qquad n>1. \tag{14}$$

The property that we will need for our later argument is

$$\lim_{\rho\to 1.0} \partial\xi/\partial\rho<\infty. \tag{15}$$

With the above definitions and assumptions we now deduce a formula of labour volume as a function of sales size. By definition, the volume of labour for a service capacity of S service channels (S attendants) during the period of length T is ST,

$$a_s = ST, \tag{16}$$

where a_s = labour volume for the service capacity. In order to express this labour volume as a function of sales size Q we proceed as follows: Substituting the definition of the utilization factor ρ [see (IX)] in (16) we find

$$a_s = \lambda\sigma T/\rho. \tag{17}$$

Substituting $\gamma_1 = \sigma/tr$ [see (12)] we then find

$$a_s = \frac{1}{\rho}\gamma_1 tr\,\lambda T. \tag{18}$$

Since tr is the average sales per customer visit and λT is the expected number of customer visits over the period, we have

$$Q = tr\,\lambda T, \tag{19}$$

where Q = the amount of sales to be expected over the period. So we now have

$$a_s = \frac{1}{\rho}\gamma_1 Q. \tag{20}$$

Let a_p be the labour volume required to perform PPA. Then, by the definition of γ_2 [see (IV)], we have

$$a_p = \gamma_2 Q. \tag{21}$$

The total labour volume over the period is less than the sum of a_s and a_p, because part of the idle service capacity can be used to perform PPA. In fact, by the definition of ξ [see (XIII)], the total labour volume is

$$a = a_s + a_p - \xi(1-\rho)a_s, \tag{22}$$

provided that

$$a_p > \xi(1-\rho)a_s,$$

which, by substitution of (20) and (21), yields

$$a=\{1-\xi(1-\rho)\}a_s+\gamma_2 Q=\{1-\xi(1-\rho)\}\frac{1}{\rho}\gamma_1 Q+\gamma_2 Q, \tag{23}$$

provided that

$$\gamma_2 Q>\xi((1-\rho)/\rho)\gamma_1 Q,$$

which is the case for sufficiently large Q.

For our proof of Theorem 1, we must express the utilization factor ρ as a function of sales size (Q) for any given value of w [held constant for all sales sizes in accordance with homogeneity assumption (X)].

With the assumptions (V) to (IX), ordinary multiple channel queuing theory yields the following relationship for the relative waiting time (w) as a function of the number of service channels (S) and the utilization factor (ρ), which we take from the literature[10] and present without proof,

$$w=S(1-\rho)\{(1-\rho)X+1\}^{-1} \text{ where } X=\sum_{i=0}^{S-1}(S!/i!)\rho^{i-S}S^{i-S}. \tag{24}$$

From this relationship we can derive the following boundary condition:

$$\lim_{S\to\infty}\rho_w=1.0 \quad \text{for any} \quad w>0. \tag{25}$$

The subscript w indicates that we are considering ρ as a function of S: we are moving along an isomene (line of equal waiting), with a constant value of w. For continuous S, a second boundary condition is

$$\lim_{S\to\infty}(\partial\rho/\partial S)_w=0. \tag{26}$$

For a further development of our model of labour costs (23), we need to express ρ as a function of S for a constant value of w. We are unable to find an analytical derivation of this function from (24), which expresses w as a function of ρ and S. Therefore we now proceed as follows:

(i) On the basis of (24) we calculate w for S ranging between 1 and 10, and for a sufficiently fine grid on ρ to find those values of ρ, for each value of S, that correspond with predetermined values of w, ranging between 0.25 and 4.0. The ranges for S and w are chosen for their plausibility

[10]Ferrero di Roccaferrera (1964, p. 840, 841).

under normal conditions of retailing. The results are given in table 1, where we tabulate ρ for the ranges chosen for S and w.[11]

(ii) We specify a class of continuous functions, expressing ρ as a function of S and w, which satisfy the boundary conditions (25) and (26).

(iii) For each value of w we estimate the coefficients of the function out of that class which gives the best least squares fit to the values tabulated in table 1.

(iv) We plot the resulting coefficients against the different values of w in order to express the coefficients as functions of w.

(v) This yields an approximation for ρ as a function of S and w (for the ranges chosen for S and w).

Table 1

ρ for different values of w and S.

	w				
S	0.25	0.50	1.0	2.0	4.0
1	0.200	0.333	0.500	0.667	0.800
2	0.447	0.577	0.707	0.816	0.894
3	0.573	0.686	0.790	0.872	0.928
4	0.648	0.747	0.835	0.901	0.945
5	0.699	0.788	0.863	0.919	0.955
6	0.736	0.816	0.883	0.932	0.963
7	0.764	0.838	0.898	0.941	0.968
8		0.854	0.909	0.947	
9		0.868	0.918	0.953	
10		0.879	0.925	0.957	

In fig. 7 we plot the values of ρ, given in table 1, against those for S, for each value of w, and we draw continuous curves connecting points of equal waiting (isomenes). The results illustrate the boundary conditions specified in (25) and (26).

Now we turn to our choice of a class of functions to approximate ρ as a function of S, for a given value of w. The following class of functions satisfies the boundary conditions (25) and (26):

$$1/\rho = 1 + d_1/S + d_2/S^2 + d_3/S^3 + \ldots + d_n/S^n, \tag{27}$$

where $d_1, d_2, \ldots, d_n$ are functions (as yet unknown) of w.

The third step now is an estimation of the coefficients $d_1, \ldots, d_n$ for different values of w, which we conduct by means of an ordinary least squares fitting procedure. The results showed an extremely close fit, for each value of w, if

[11]The step function in fig. 3 was constructed on the basis of this table.

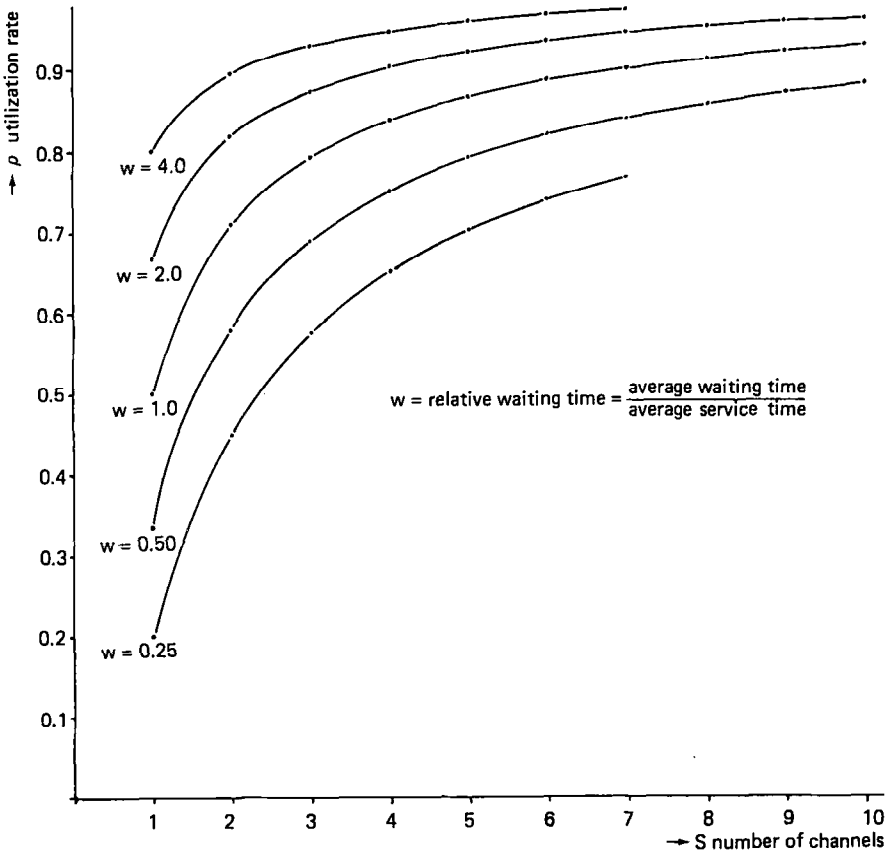

Fig. 7. Isomenes.

we truncated after d_3/S^3, and they showed that the coefficient d_2 then was practically zero. Thus our approximation now is

$$\frac{1}{\rho}=1+d_1/S+d_3/S^3. \tag{28}$$

The estimates of d_1 and d_3 are given in table 2.

d_1 and d_3 are functions of w, and the fourth step in our procedure is to approximate those functions.

In fig. 8 we plot d_1 and d_3 against w on double-log paper. The figure shows that for the range of w considered the relationships are approximately log-linear:

Table 2
Estimates of d_1 and d_3 for different values of w.[a]

w	d_1	d_3
0.25	2.020	1.98
0.50	1.302	0.697
1.00	0.776	0.224
2.00	0.434	0.066
4.00	0.232	0.018

[a]The results were obtained with an OLS regression of $1/\rho - 1$ on $1/S$ and $1/S^3$. The resulting fit was so close that the deviations could not be made to show in a graphical presentation, for the ranges of S and w considered (cf. fig. 7).

$$\log d_1 = \log \delta_{10} + \delta_{11} \log w, \qquad \log d_3 = \log \delta_{30} + \delta_{31} \log w. \tag{29}$$

We estimate δ_{11} and δ_{31} by reading off the slopes of the linear approximations in fig. 8, and we estimate δ_{10} and δ_{30} by reading off the values of d_1 and d_3 for $w = 1.0$. The results are

$$\delta_{11} \approx -5/6, \qquad \delta_{31} \approx -5/3,$$

$$\delta_{10} \approx 0.775, \qquad \delta_{30} \approx 0.225.$$

Thus our end result is as follows:

$$\frac{1}{\rho} = 1 + 0.775(1/w)^{5/6}/S + 0.225(1/w)^{5/3}/S^3. \tag{30}$$

With this result we can now prove that the cost function specified in (23) has the properties stated in Theorem 1 [see (10), (11) and (12)]. The proof is given in the appendix.

5. Evaluation and further research

We claim that our theory yields a promising 'research programme',[12] which:

[12]Cf. Lakatos in Lakatos (1976).

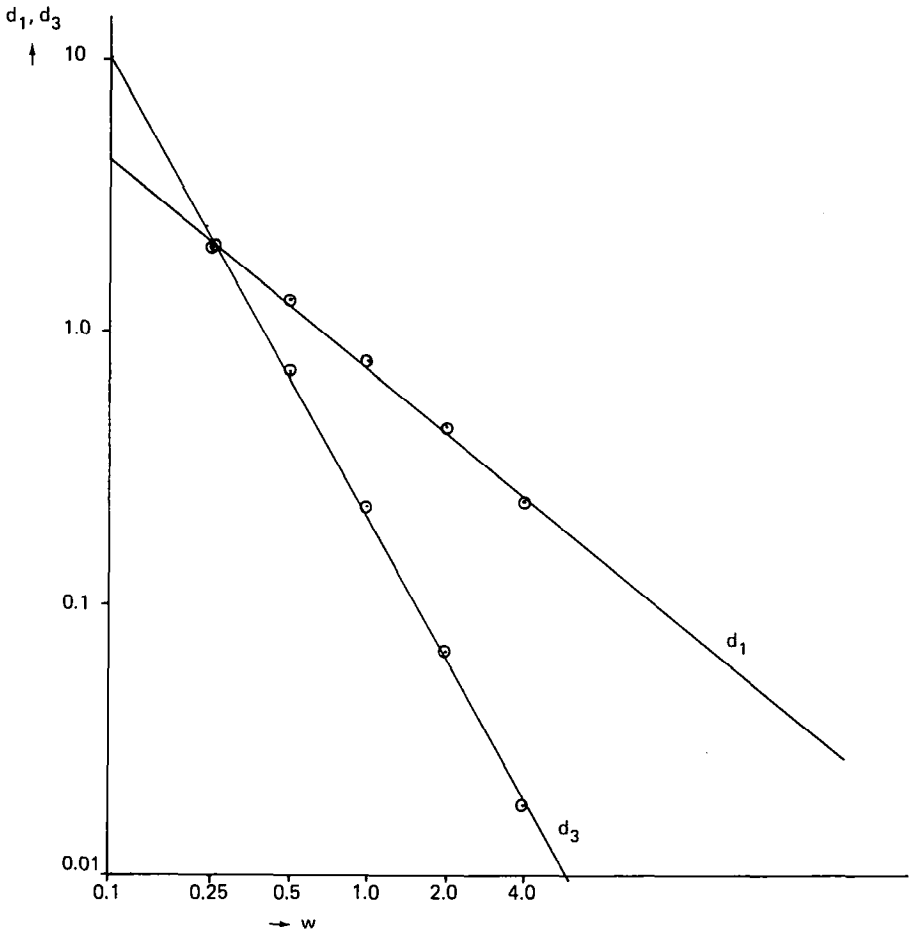

Fig. 8. Log d_1 and log d_3 against log w.

— is based on concepts and explanatory principles that are plausible to people who know the retailing sector;
— has not been falsified: there has been much corroboration, and so far the empirical anomalies that did appear could later be explained from within the theory, in a manner which was not purely ad hoc, and were actually turned into confirming evidence;
— yields accurately predicted 'novel facts';
— leads to useful applications;
— provides a fruitful basis for ongoing research.

In the attempt to obtain observations on a set of shops that are more

homogeneous than the independent shopkeepers from which we induced our theory, in order to obtain a narrower scatter to test the linear cost curve, we obtained data on the shops of a large Dutch chain store enterprise. The results are illustrated in fig. 9. In other studies apparent falsifications of the linear cost curve were later explained by inhomogeneities, notably with respect to assortment composition and the number of departments (in correlation with sales size), according to which the theory predicted deviations in the directions in which they were in fact found.

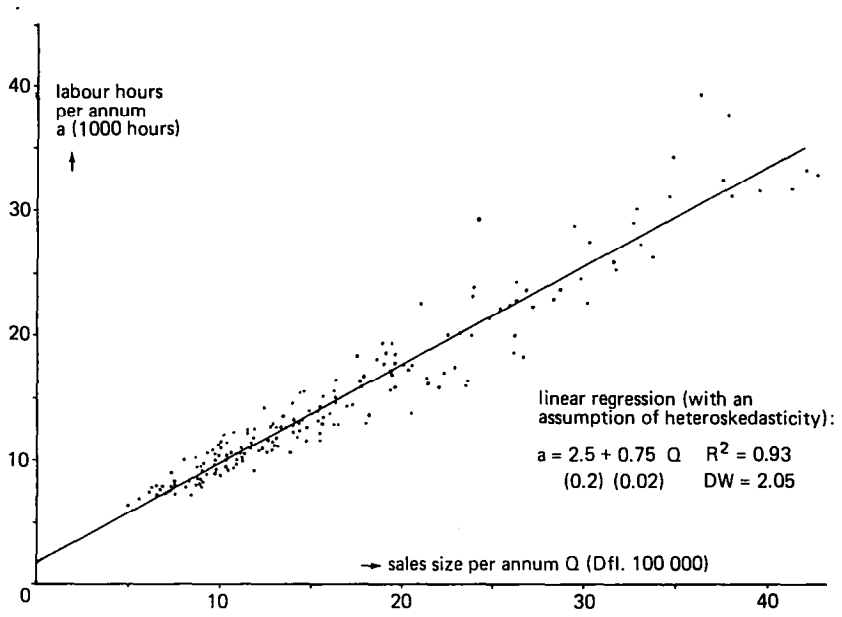

Fig. 9a. Chain store self-service grocers, 1974.

Some more specific, quantitative results were the following:

(i) In a study of the supermarkets of the Dutch chain store enterprise our empirical estimates of threshold labour were 13.4 thousand hours, with standard error 2.0, for 1973, and 9.7 thousand hours, with standard error 1.1, for 1974. On the basis of an annual opening time of 2.5 thousand hours, we inferred on the basis of our theory that the average number of departments was about 5, with, in view of the standard errors, a certain variation of the number of departments across individual shops. Subsequently, this was confirmed by the management of the enterprise.[13] Management also

[13]Typically, a supermarket had, in addition to the basic self-service section for groceries, service points for fresh meats, fresh vegatables/fruit, delicatessen and fresh bakery products.

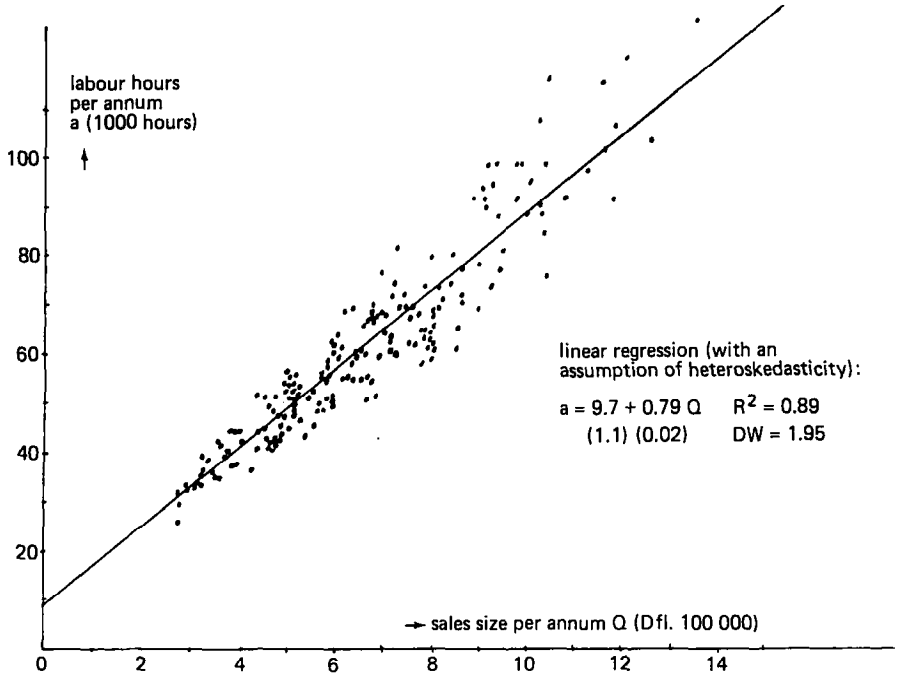

Fig. 9b. Chain store supermarkets, 1974.

confirmed the underlying assumption of an average waiting time which is, on the whole, about equal to the average service time: the policy of the enterprise was to add to the labour capacity of a service point, during a certain time interval, if the queue length was consistently larger than 2.0.

(ii) In a study of independent grocery stores in the U.S. in 1948[14] our information indicated an annual opening time of about 4000 hours (due to evening and week-end openings, which are normal in the U.S. grocery trade),[15] and a man-year of about 2000 hours.[16] Thus we predicted a threshold labour of 2.0 persons engaged (in full-time equivalents).[17] The empirical estimate was 1.97 with standard error 0.10.[18] For independent clothes shops in the U.S. the annual opening time was, according to our information, about 2900 hours (with some evening openings but closure on

[14] *Source of the data:* Hall, Knapp and Winsten (1961).

[15] *Source:* Kruiderink (1966).

[16] In the U.S. the 5-day, 40-hour labour week was introduced as early as 1938 (*source:* U.S. Embassy, The Hague).

[17] On the 'reasonable' assumption that at that early date (1948) the vast majority of the (relative small) independent grocers still had only one department (no additional specialized service counters).

[18] The estimates were conducted on the basis of regional averages per shop.

Sundays), yielding a prediction of 1.45 persons engaged for threshold labour. The empirical estimate was 1.57 with standard error 0.27.

In Canada, annual opening times of grocers were lower than in the U.S., due to a much stronger influence of the churches, yielding a prohibition of Sunday openings in most provinces. Our prediction of threshold labour was 1.4 persons engaged, and the empirical estimate was 1.31 with standard error 0.25.

For British shops in 1950, the labour week was still, according to our information, about 46 hours,[19] yielding a prediction of 1.08 for threshold labour, and the empirical estimate was 1.04 with standard error 0.24.

In practice no set of observations will satisfy the stringent assumptions of homogeneity that we made in the previous section, and extensions must be made to take into account differences in: assortment composition; the use of part-time labour; the cost, quality and availability of labour; waiting times and service times in relation to the local intensity of competition; patterns of the fluctuation of demand in relation to the type of location; average sales per customer visit in relation to the type of customer and type of location; the services rendered by wholesalers, manufacturers and the central staff of (voluntary) chains; the use of technical devices and methods for cashing, checking, ordering, administration and product handling; etc.

Apart from the effect of differences in assortment composition, the provision of service (as opposed to self-service) and own production, deviations from the average linear cost curve were indeed to some extent explained, empirically, by differences in the use of part-time labour, the wage rate (with an elasticity of about 0.6), transaction size and type of location.[20]

There is scope for mathematical refinements with respect to queue discipline; an expected 'survival rate' of impatient customers as a function of queue length; the proportion of idle service capacity that can be used for product handling on the basis of a frequency distribution of the duration of idle periods; a service time which is not proportional to transaction size; etc. There is scope for econometric refinements with respect to stochastic structure of disturbances, estimation procedures and tests of linearity. Several of these further developments are under way.

Appendix: Proof of Theorem 1

For a proof of Theorem 1, we must show that:

— the cost function (23) has an asymptote $y=\beta_0+\beta_1 Q$;
— $w=0.75\rightarrow\beta_0=T$ (T is the length of the period considered);
— $\beta_1=\gamma_1+\gamma_2$.

[19] *Source:* Hall, Knapp and Winsten (1961).
[20] Cf. Nooteboom (1980, chs. 5 and 6).

The asymptote, if there is any, is found by taking a tangent to the cost curve (23) at any point p (with coordinates a_p, Q_p) and then taking its limit for $Q_p \to \infty$.

Let the tangent at point p be expressed as

$$y = \varepsilon_0 + \varepsilon_1 Q. \tag{A.1}$$

Then we must have

$$\varepsilon_1 = \left(\frac{\partial a}{\partial Q}\right)_p \quad \text{and} \quad \varepsilon_0 = a_p - \left(\frac{\partial a}{\partial Q}\right)_p Q_p. \tag{A.2}$$

Then the asymptote, if there is any, is defined as

$$y = \beta_0 + \beta_1 Q, \qquad \beta_0 = \lim_{Q_p \to \infty} \varepsilon_0, \quad \beta_1 = \lim_{Q_p \to \infty} \varepsilon_1. \tag{A.3}$$

From now on we drop the subscript p.

Both ε_0 and ε_1 are functions of $\partial a/\partial Q$. From (23) we find

$$\frac{\partial a}{\partial Q} = \{1 - \xi(1-\rho)\}\frac{\partial a_s}{\partial Q} + a_s \frac{\partial \rho}{\partial Q}\left\{\xi - (1-\rho)\frac{\partial \xi}{\partial \rho}\right\} + \gamma_2. \tag{A.4}$$

In order to proceed, we need to specify $\partial a_s/\partial Q$ as a function of Q.

Lemma 1

$$\frac{\partial a_s}{\partial Q} = \frac{a_s/Q}{1 + (S/\rho)(\partial \rho/\partial S)}. \tag{A.5}$$

Proof. According to an earlier result we had $a_s = (1/\rho)\gamma_1 Q$ [see (20)]; differentiation with respect to Q yields

$$a_s \frac{\partial \rho}{\partial Q} + \rho \frac{\partial a_s}{\partial Q} = \gamma_1, \tag{A.6}$$

by definition we had $a_s = ST$ [see (16)], using this, we find

$$\frac{\partial \rho}{\partial Q} = \frac{\partial \rho}{\partial S}\frac{\partial S}{\partial Q} = \frac{\partial \rho}{\partial S}\frac{\partial a_s}{\partial Q}\frac{1}{T}, \tag{A.7}$$

substituting (A.7) and $a_s = ST$ in (A.6), we find (A.5). QED

On the basis of Lemma 1 we obtain for ε_0:

Lemma 2

$$\varepsilon_0 = \left\{1-\xi+\rho(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}\frac{(S^2/\rho)(\partial\rho/\partial S)T}{1+(S/\rho)(\partial\rho/\partial S)}. \tag{A.8}$$

Proof. Substitution of (A.4) in (A.2) followed by substitution of (A.7) yields

$$\varepsilon_0 = \{1-\xi(1-\rho)\}a_s - \left[1-\xi(1-\rho)+S\frac{\partial\rho}{\partial S}\left\{\xi-(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}\right]\frac{\partial a_S}{\partial Q}Q, \tag{A.9}$$

substitution of (A.5) in (A.9) yields

$$\varepsilon_0 = \left\{1-\xi+\rho(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}\frac{(S/\rho)(\partial\rho/\partial S)a_s}{1+(S/\rho)(\partial\rho/\partial S)}. \tag{A.10}$$

substitution of $a_s = ST$ in (A.10) yields (A.8). QED

On the basis of Lemma 1 we obtain for ε_1:

Lemma 3

$$\varepsilon_1 = \left[\frac{1-\xi}{\rho}+\xi+\frac{S}{\rho}\frac{\partial\rho}{\partial S}\left\{\xi-(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}\right]\frac{\gamma_1}{1+(S/\rho)(\partial\rho/\partial S)}+\gamma_2. \tag{A.11}$$

Proof. Substitution of (A.4) in (A.2) followed by substitution of (A.7) yields

$$\varepsilon_1 = \{1-\xi(1-\rho)\}\frac{\partial a_s}{\partial Q}+a_s\frac{\partial\rho}{\partial S}\frac{\partial a_s}{\partial Q}\frac{1}{T}\left\{\xi-(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}+\gamma_2, \tag{A.12}$$

substitution of $a_s = ST$ in (A.12) followed by substitution of (A.5) yields

$$\varepsilon_1 = \left[1-\xi(1-\rho)+S\frac{\partial\rho}{\partial S}\left\{\xi-(1-\rho)\frac{\partial\xi}{\partial\rho}\right\}\right]\frac{a_s/Q}{1+(S/\rho)(\partial\rho/\partial S)}+\gamma_2, \tag{A.13}$$

substitution of $a_s = (1/\rho)\gamma_1 Q$ [see (20)] in (A.13) yields (A.11). QED

Now we need to take the limits of ε_0 and ε_1 for $Q\to\infty$. To do so, we need

to know the limits for $Q \to \infty$ of $\partial\xi/\partial\rho$, of $(S/\rho)(\partial\rho/\partial S)$ and of $(S^2/\rho)(\partial\rho/\partial S)$. According to the model assumed for the relationship between ξ and ρ we had [see (15)]

$$\lim_{\rho \to 1} \partial\xi/\partial\rho < \infty. \tag{A.14}$$

Our model for the relationship between ρ and S was [see (27)]

$$\frac{1}{\rho} = 1 + d_1/S + d_2/S^2 + \ldots + d_n/S^n. \tag{A.15}$$

From this it follows that

$$\partial\rho/\partial S = \rho^2 (d_1/S^2 + 2d_2/S^3 + \ldots + n d_n/S^{n+1}), \tag{A.16}$$

so that

$$\lim_{S \to \infty} (S/\rho)(\partial\rho/\partial S) = 0 \quad \text{and} \quad \lim_{S \to \infty} (S^2/\rho)(\partial\rho/\partial S) = d_1. \tag{A.17}$$

In our approximation, for a given range of S and w, the value of d_1 was [see (30)]:

$$d_1 = 0.775\,(1/w)^{5/6}. \tag{A.18}$$

For any constant value of the relative waiting time w,

$$Q \to \infty \Rightarrow S \to \infty \Rightarrow \rho \to 1.0 \Rightarrow \xi \to 0. \tag{A.19}$$

We now use the results to deduce the limits of the ε_0 and ε_1 specified in Lemmas 2 and 3 [(A.8) and (A.11)]. The results are as follows:

$$\beta_0 = \lim_{Q \to \infty} \varepsilon_0 = d_1 T = 0.775\,(1/w)^{5/6} T \tag{A.20}$$

(for the ranges of S and w to which the approximation applies), and

$$\beta_1 = \lim_{Q \to \infty} \varepsilon_1 = \gamma_1 + \gamma_2. \tag{A.21}$$

(A.20) yields

$$\beta_0 = T \quad \text{for} \quad w \approx 0.75 \tag{A.22}$$

(more precisely: $w = 0.736$). This completes the proof of Theorem 1.

For the cost curve on an annual basis, we segment the year into different subperiods with different expected rates of customer arrivals. Under the assumption that use is made of part-time labour to adapt the service capacity to the different intensities of demand in the different subperiods, we arrive at the annual cost function by an aggregation of the cost functions per period. For the aggregate cost function the conclusions (A.20), (A.21) and (A.22) also apply, where T then is the annual opening time α_0. The procedure is straightforward, and we will not present the proof here.[21]

[21]For the proof we refer to Nooteboom (1980, app. 3.3).

References

Arndt, J. and Olsen, 1975, A research note on economies of scale in retailing, Swedish Journal of Economics.
McClelland, W.G.M., 1966, Costs and competition in retailing (New York).
Ferrero di Roccaferrera, G.M., 1964, Operations research models for business and industry (Cincinnati).
Hall, M., J. Knapp and G. Winsten, 1961, Distribution in Great Britain and North America (Oxford).
Holdren, B.R., 1960, The structure of a retail market and the market behaviour of retail units (Englewood Cliffs).
Kruiderink, W.H., 1966a, Tussen Atlantic en Pacific (Nijmegen).
Kruiderink, W.H., 1966b, Van Washington tot Montreal (Nijmegen).
Lakatos, I., ed., 1976, Criticism and the growth of knowledge (Cambridge).
Nooteboom, B., 1980, Retailing: Applied analysis in the theory of the firm (Amsterdam/Uithoorn).
Palamountain, J.C., 1955, The politics of distribution (Cambridge, MA).
Pickering, J.F., 1972, Economic implications of hypermarkets in Britain, European Journal of Marketing.
Twentieth Century Fund, 1939, Does distribution cost too much? (New York).